Year	Sem.	Subject Code	Title of the paper	Hours/ Week
2018 -2019 onwards	VI	18BBO63C	CORE PAPER XII BIOTECHNOLOGY AND BIOINFORMATICS	5

Unit - IV

Bioinformatics: Fundamentals of computer hardware components, software types- operating system software (Windows and UNIX) and programme software (BioPEARL), Internet browser, HTML, Databases, Data mining and Data retrieval.

Prepared by
Dr. M M Sudheer Mohammed
Associate Professor of otany
Mobile-9443274469

Introduction to Computer

Computer

A computer is an electronic device, operating under the control of instructions stored in its own memory that can accept data (input), process the data according to specified rules, produce information (output), and store the information for future use[1].

Functionalities of a computer

Any digital computer carries out five functions in gross terms:

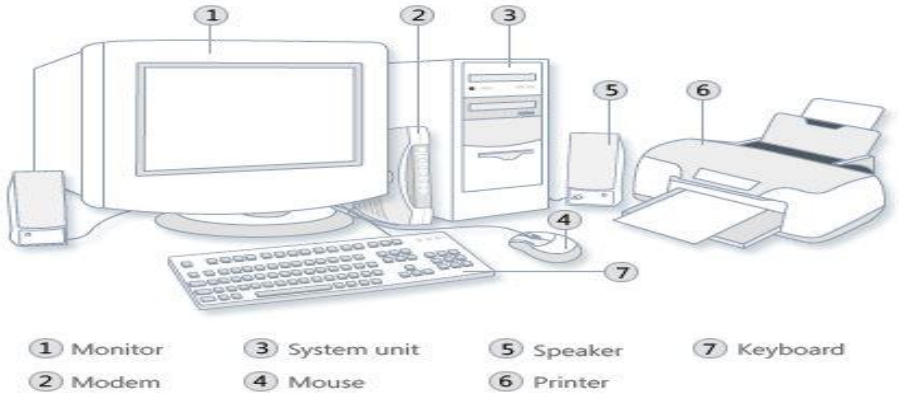
- Takes data as input.
- Stores the data/instructions in its memory and use them when required.
- Processes the data and converts it into useful information.
- Generates the output
- Controls all the above four steps.

Computer Components

Any kind of computers consists of **HARDWARE** AND **SOFTWARE**.

Hardware:

Computer hardware is the collection of physical elements that constitutes a computer system. Computer hardware refers to the physical parts or components of a computer such as the monitor, mouse, keyboard, computer data storage, hard drive disk (HDD), system unit (graphic cards, sound cards, memory, motherboard and chips), etc. all of which are physical objects that can be touched.



Input Devices

Input device is any peripheral (piece of computer hardware equipment to provide data and control signals to an information processing system such as a computer or other information appliance.

Input device Translate data from **form** that humans understand to one that the computer can work with. Most common are keyboard and mouse

Example of Input Devices:-

1. Keyboard	2. Mouse (pointing device)	3. Microphone
4. Touch screen	5. Scanner	6. Webcam
7. Touchpads	8. MIDI keyboard	9.
10.Graphics Tablets	11.Cameras	12.Pen Input
13.Video Capture Hardware	14.Microphone	15.Trackballs
16.Barcode reader	17.Digital camera	18.Joystick
19.Gamepad	20.Electronic Whiteboard	21.

Note: The most common use keyboard is the QWERTY keyboard. Generally standard Keyboard has 104 keys.

Central Processing Unit (CPU)

A CPU is brain of a computer. It is responsible for all functions and processes. Regarding computing power, the CPU is the most important element of a computer system.

The CPU is comprised of three main parts :

* ***Arithmetic Logic Unit (ALU)***: Executes all arithmetic and logical operations. Arithmetic calculations like as addition, subtraction, multiplication and division. Logical operation like compare numbers, letters, or special characters
* ***Control Unit (CU):*** controls and co-ordinates computer components.
    1. Read the code for the next instruction to be executed.
    2. Increment the program counter so it points to the next instruction.
    3. Read whatever data the instruction requires from cells in memory.
    4. Provide the necessary data to an ALU or register.
    5. If the instruction requires an ALU or specialized hardware to complete, instruct the hardware to perform the requested operation.
* ***Registers*** :Stores the data that is to be executed next, "very fast storage area".

Primary Memory:-

1. **RAM**: Random Access Memory (RAM) is a memory scheme within the computer system responsible for storing data on a temporary basis, so that it can be promptly accessed by the processor as and when needed. It is volatile in nature, which means that data will be erased once supply to the storage device is turned off. RAM stores data randomly and the processor accesses these data randomly from the RAM storage. RAM is considered "random access" because you can access any memory cell directly if you know the row and column that intersect at that cell.
2. **ROM** (Read Only Memory): ROM is a permanent form of storage. ROM stays active regardless of whether power supply to it is turned on or off. ROM devices do not allow data stored on them to be modified.

Secondary Memory:-

Stores data and programs permanently :its retained after the power is turned off

1. Hard drive (HD): A hard disk is part of a unit, often called a "disk drive," "hard drive," or "hard disk drive," that store and provides relatively quick access to large amounts of data on an electromagnetically charged surface or set of surfaces.
2. Optical Disk: an optical disc drive (ODD) is a disk drive that uses laser light as part of the process of reading or writing data to or from optical discs. Some drives can only read from discs, but recent drives are commonly both readers and recorders, also called burners or writers. Compact discs, DVDs, and Blu-ray discs are common types of optical media which can be read and recorded by such drives. Optical drive is the generic name; drives are usually described as "CD" "DVD", or "Bluray", followed by "drive", "writer", etc. There are three main types of optical media: CD, DVD, and Blu-ray disc. CDs can store up to 700 megabytes (MB) of data and DVDs can store up to 8.4 GB of data. Blu-ray discs, which are the newest type of optical media, can store up to 50 GB of data. This storage capacity is a clear advantage over the floppy disk storage media (a magnetic media), which only has a capacity of 1.44 MB.

3. Flash Disk

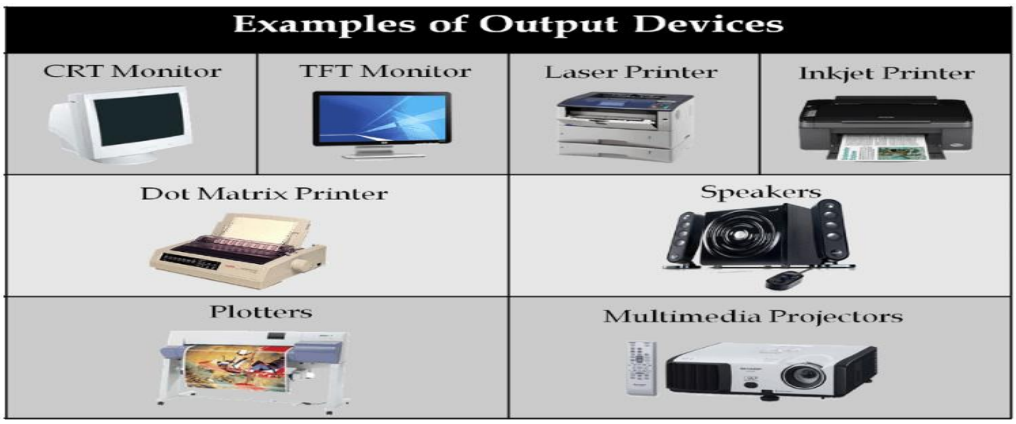
A storage module made of flash memory chips. A Flash disks have no mechanical platters or access arms, but the term "disk" is used because the data are accessed as if they were on a hard drive. The disk storage structure is emulated.

Comparison between Main memory (RAM) and Secondary Memory (Hard disk)

RAM	Hard Disk (Hard Drive)
Memory	Storage
Smaller amount (typically 500 MB-6 GB)	Much larger amount (typically 80GB to 1000 GB)
Temporary storage of files and programs	Permanent storage of files and programs
A little like your real desktop - has only your current work on it (which could be ruined by a spill of Coke or coffee!)	Like a file cabinet - has long-term storage of work (it's safe from spills!)
Contents disappear when you turn off power to the computer and when the computer crashes	Contents remain when you turn off the power to the computer (they don't disappear unless you purposely delete them), and when the computer crashes
Consists of chips (microprocessors)	Consists of hard disks (platters)
When you want to use a program, a temporary copy is put into RAM and that's the copy you use	Holds the original copy of the program permanently

Output devices

An output device is any piece of computer hardware equipment used to communicate the results of data processing carried out by an information processing system (such as a computer) which converts the electronically generated information into human-readable form.



Example on Output Devices:

1. Monitor	2. LCD Projection Panels
3. Printers (all types)	4. Computer Output Microfilm (COM)
5. Plotters	6. Speaker(s)
7. Projector	

Note Basic types of monitors are a.Cathode Ray Tube (CRT). B. Liquid Crystal Displays (LCD). c.light-emitting diode (LED).
Printer types: 1-Laser Printer. 2-Ink Jet Printer. 3-Dot Matrix Printer

Software

Software is a generic term for organized collections of computer data and instructions, often broken into two major categories: system software that provides the basic non-task-specific functions of the computer, and application software which is used by users to accomplish specific tasks.

Software Types

A. System software is responsible for controlling, integrating, and managing the individual hardware components of a computer system so that other software and the users of the system see it as a functional unit without having to be concerned with the low-level details such as transferring data from memory to disk, or rendering text onto a display. Generally, system software consists of an operating system and some fundamental utilities such as disk formatters, file managers, display managers, text editors, user authentication (login) and management tools, and networking and device control software.

B. **Application software** is used to accomplish specific tasks other than just running the computer system. Application software may consist of a single program, such as an image viewer; a small collection of programs (often called a software package) that work closely together to accomplish a task, such as a spreadsheet or text processing system; a larger collection (often called a software suite) of related but independent programs and packages that have a common user interface or shared data format, such as Microsoft Office, which consists of closely integrated word processor, spreadsheet, database, etc.; or a software system, such as a database management system, which is a collection of fundamental programs that may provide some service to a variety of other independent applications.

Comparison Application Software and System Software

	System Software	Application Software
	Computer software, or just software is a general term primarily used for digitally stored data such as computer programs and other kinds of information read and written by computers. App comes under computer software though it has a wide scope now.	Application software, also known as an application or an "app", is computer software designed to help the user to perform specific tasks.
Example:	1) Microsoft Windows 2) Linux 3) Unix 4) Mac OSX 5) DOS	1) Opera (Web Browser) 2) Microsoft Word (Word Processing) 3) Microsoft Excel (Spreadsheet software) 4) MySQL (Database Software) 5) Microsoft PowerPoint (Presentation Software) 6) Adobe Photoshop (Graphics Software)
Interaction:	Generally, users do not interact with system software as it works in the background.	Users always interact with application software while doing different activities.
Dependency:	System software can run independently of the application software.	Application software cannot run without the presence of the system software.

Unit of Measurements

Storage measurements: The basic unit used in computer data storage is called a bit (binary digit). Computers use these little bits, which are composed of ones and zeros, to do things and talk to other computers. All your files, for instance, are kept in the computer as binary files and translated into words and pictures by the software (which is also ones and zeros). This two number system, is called a "binary number system" since it has only two numbers in it. The decimal number system in contrast has ten unique digits, zero through nine.

Computer Storage units

Bit	BIT	0 or 1
Kilobyte	KB	1024 bytes
Megabyte	MB	1024 kilobytes
Gigabyte	GB	1024 megabytes
Terabyte	TB	1024 gigabytes

Size example

- 1 bit - answer to an yes/no question
- 1 byte - a number from 0 to 255.
- 90 bytes: enough to store a typical line of text from a book.
- 4 KB: about one page of text.
- 120 KB: the text of a typical pocket book.
- 3 MB - a three minute song (128k bitrate)
- 650-900 MB - an CD-ROM
- 1 GB -114 minutes of uncompressed CD-quality audio at 1.4 Mbit/s
- 8-16 GB - size of a normal flash drive

Speed measurement: The speed of Central Processing Unit (CPU) is measured by Hertz (Hz), Which represent a CPU cycle. The speed of CPU is known as Computer Speed.

CPU SPEED MEASURES	
1 hertz or Hz	1 cycle per second
1 MHz	1 million cycles per second or 1000 Hz
1 GHz	1 billion cycles per second or 1000 MHz

Computers classification***

Computers can be generally classified by size and power as follows, though there is Considerable overlap:

- Personal computer: A small, single-user computer based on a microprocessor. In addition to the microprocessor, a personal computer has a keyboard for entering data, a monitor for displaying information, and a storage device for saving data.
- workstation : A powerful, single-user computer. A workstation is like a personal computer, but it has a more powerful microprocessor and a higher-quality monitor.
- minicomputer : A multi-user computer capable of supporting from 10 to hundreds of users simultaneously.
- mainframe : A powerful multi-user computer capable of supporting many hundreds or thousands of users simultaneously.
- supercomputer : An extremely fast computer that can perform hundreds of millions of instructions per second.

Laptop and Smartphone Computers

LAPTOP: A laptop is a battery or AC-powered personal computer that can be easily carried and used in a variety of locations. Many laptops are designed to have all of the functionality of a desktop computer, whichmeans they can generally run the same software and open the same types of files. However, some laptops, such as netbooks, sacrifice some functionality in order to be even more portable.

Netbook: A netbook is a type of laptop that is designed to be even more portable. Netbooks are often cheaper than laptops or desktops. They are generally less powerful than other types of computers, but they provide enough power for email and internet access, which is where the name "netbook" comes from.

Mobile Device: A mobile device is basically any handheld computer. It is designed to be extremely portable, often fitting in the palm of your hand or in your pocket. Some mobile devices are more powerful, and they allow you to do many of

***http://www.acobas.net/teaching/survival/handouts/pcwebopedia.pdf

the same things you can do with a desktop or laptop computer. These include tablet computers, e-readers, and smartphones.

Tablet Computers: Like laptops, tablet computers are designed to be portable. However, they provide a very different computing experience. The most obvious difference is that tablet computers don't have keyboards or touchpads. Instead, the entire screen is touch-sensitive, allowing you to type on a virtual keyboard and use your finger as a mouse pointer. Tablet computers are mostly designed for consuming media, and they are optimized for tasks like web browsing, watching videos, reading e-books, and playing games. For many people, a "regular" computer like a desktop or laptop is still needed in order to use some programs. However, the convenience of a tablet computer means that it may be ideal as a second computer.

Smartphones: A smartphone is a powerful mobile phone that is designed to run a variety of applications in addition to phone service. They are basically small tablet computers, and they can be used for web browsing, watching videos, reading e-books, playing games and more.

Data, Information and Knowledge

Data: Facts and figures which relay something specific, but which are not organized in any way and which provide no further information regarding patterns, context, etc. So data means "unstructured facts and figures that have the least impact on the typical manager."

Information: For data to become information, it must be contextualized, categorized, calculated and condensed. Information thus paints a bigger picture; it is data with relevance and purpose. It may convey a trend in the environment, or perhaps indicate a pattern of sales for a given period of time. Essentially information is found "in answers to questions that begin with such words as who, what, where, when, and how many".

Knowledge: Knowledge is closely linked to doing and implies know-how and understanding. The knowledge possessed by each individual is a product of his experience, and encompasses the norms by which he evaluates new inputs from his surroundings.

Bioperl

Bioperl, perhaps the oldest of the Bio projects, is a group of more than 500 Perl modules having numerous bioinformatics utilities and have been written and maintained by an international group of volunteers. The bioperl-live repository contains the core functionality and additional packages are for creating graphical interfaces (bioperl-gui), setting up persistent ORM storage in RDMBS (bioperl-db), running and parsing the results from hundreds of bioinformatics applications (bioperl-run), and software to automate bioinformatics analyses (bioperl pipeline).

It also has data models and operations for ontologies, phylogenetic trees, genetic maps and markers and population genetics.

- a set of Perl modules for manipulating genomic and other biological data
- an open source software project
- developed by a loose collaboration of individuals worldwide, similar to the open development ethos in Linux or Apache
- played an integral role in the Human Genome Project
- available at http://bioperl.org
- Read in sequence data from a file in standard formats (FASTA, GenBank, EMBL, SwissProt,...)
- Manipulate sequences, reverse complement, translate coding DNA sequence to protein.
- Parse a BLAST report, get access to every bit of data in the report.

The Bioperl project is an international open-source collaboration of biologists, bioinformaticians, and computer scientists that has evolved over the past 7 yrs into the most comprehensive library of Perl modules available for managing and manipulating life-science information. Bioperl provides an easy-to-use, stable, and consistent programming interface for bioinformatics application programmers. The Bioperl modules have been successfully and repeatedly used to reduce otherwise complex tasks to only a few lines of code. The Bioperl object model has been proven to be flexible enough to support enterprise-level applications such as EnsEMBL, while maintaining an easy learning curve for novice Perl programmers. Bioperl is capable of executing analyses and processing results from programs such as BLAST, ClustalW, or the EMBOSS suite. Interoperation with modules written in Python and Java is supported through the evolving BioCORBA bridge. Bioperl provides access to data stores such as GenBank and SwissProt via a flexible series of sequence input/output modules, and to the emerging common sequence data storage format of the Open Bioinformatics Database Access project.

BioJava

BioJava is a mature open-source project that provides a framework for processing of biological data. BioJava contains powerful analysis and statistical routines, tools for parsing common file formats and packages for manipulating sequences and 3D structures. It enables rapid bioinformatics Application development in the Java programming language.

Availability: BioJava is an open-source project distributed under the Lesser GPL (LGPL). BioJava can be downloaded from the BioJava website (http://www.biojava.org). BioJava requires Java 1.5 or higher.

BioJava was conceived in 1999 by Thomas Down and Matthew Pocock as an Application Programming Interface (API) to simplify bioinformatics software development using Java. It has since then evolved to become a fully featured framework with modules for performing many common bioinformatics tasks. The goal of BioJava is to facilitate code reuse and to provide standard implementations that are easy to link to external scripts and applications.

BioJava contains a number of mature APIs. The 10 most frequently used are: (1) nucleotide and amino acid alphabets, (2) BLAST parser, (3) sequence I/O, (4) dynamic programming, (5) structure I/O and manipulation, (6) sequence manipulation, (7) genetic algorithms, (8) statistical distributions, (9) graphical user interfaces and (10) serialization to databases. Below follows a short discussion of some of these modules. At the core of BioJava is a symbolic alphabetAPI which represents sequences as a list of references to singleton symbol objects that are derived from an alphabet. Lists of symbols are stored whenever possible in a compressed form of up to four symbols per byte of memory. In addition to the fundamental symbols of a given alphabet (A, C, G and T in the case of DNA), all BioJava alphabets implicitly contain extra symbol objects representing all possible combinations of the fundamental symbols.

BioJava is a mature project and has been employed in a number of real-world applications and over 50 published studies. A list of these can be found on the BioJava website. According to the project tracking web site Ohloh (http://www.ohloh.net/projects/biojava), the BioJava code-base represents an estimated 47 person-years' worth of effort.

Internet browser

A web browser (commonly referred to as a browser) is a software application for retrieving, presenting, and traversing information resources on the World Wide Web (also known as the internet or the Net). The most popular web browsers are Google Chrome, Microsoft Edge (formerly Internet Explorer), Mozilla Firefox, and Apple's Safari. If you have a Windows computer, Microsoft Edge (or its older counterpart, Internet Explorer) are already installed on your computer. If you are running an Apple computer, you already have Safari installed on your computer. You may also have other browsers installed on your computer. If the browser you want to use is not installed on the computer, download links for Chrome and Firefox are described below.

There are three primary browsers that are used to access the internet that you need to be able to identify and use. Each of these browsers is made by a separate company, and has a different look, but there are many tools and shortcuts that can generally be used on any of the browsers.

Microsoft Edge

You have most likely heard the name of one or more of these browsers, but let's first identify the Microsoft Edge browser. Unlike other browsers, which must be downloaded, Microsoft Edge (or Internet Explorer) comes with Windows. That means if you have a Windows computer, Edge (or Explorer) is already on your computer. The Edge icon on a Windows 10 computer system can be found either on the bottom taskbar or along the side. Click on the icon with the mouse and it will open the browser. The icon might be in slightly different places on your desktop, but look for the icon and double click on it to open the browser. Regardless of which version of Windows you have, you can also open the browser from the start menu. Select the start button, and on clicking the icon for Edge, it will open.

Google Chrome

One of the most popular web browsers is Google Chrome (often simply referred to as Chrome). This icon is associated with the Chrome browser. This browser will be used as an example in this course in order to learn basic browser navigation skills and perform the other assignments and tasks associated with this module.

If Chrome is on your computer, the Chrome browser icon can be found on the Windows 10 desktop in the bottom task bar or along the side. Regardless of which version of Windows you have, you can also open the browser from the start menu. Select the start button and type in Chrome. If the

Chrome browser is on your computer, it will be displayed in the menu, where you can now see the icon and select it to open.

Mozilla Firefox

Mozilla Firefox (often simply referred to as Firefox) is a browser created by the company Mozilla and is another browser frequently used to "surf" or search the World Wide Web. This is the icon associated with the Firefox browser.If Firefox is on your computer, the Firefox icon on your Windows 10 computer system can be found either on the bottom taskbar or along the side. The icon might be in slightly different places on your desktop, but look for the icon and click on it to open the browser. Just like the other two browsers, it may also be opened in the Start menu box by typing Firefox, then selecting the option to open it.

HTML: HyperText Markup Language

HTML (HyperText Markup Language) is the most basic building block of the Web. It defines the meaning and structure of web content. Other technologies besides HTML are generally used to describe a web page's appearance/presentation (CSS) or functionality/behavior (JavaScript). "Hypertext" refers to links that connect web pages to one another, either within a single website or between websites. Links are a fundamental aspect of the Web. By uploading content to the Internet and linking it to pages created by other people, you become an active participant in the World Wide Web.

HTML uses "markup" to annotate text, images, and other content for display in a Web browser. HTML markup includes special "elements" such as <head>, <title>, <body>, <header>, <footer>, <article>, <section>, <p>, , , <img>, <aside>, <audio>, <canvas>, <datalist>, <details>, <embed>, <nav>, <output>, <progress>, <video>, <ul>, <ol>, <li> and many others. An HTML element is set off from other text in a document by "tags", which consist of the element name surrounded by "<" and ">". The name of an element inside a tag is case insensitive. That is, it can be written in uppercase, lowercase, or a mixture. For example, the <title> tag can be written as <Title>, <TITLE>, or in any other way.

HTML is a MUST for students and working professionals to become a great Software Engineer specially when they are working in Web Development Domain.

i) Create Web site - You can create a website or customize an existing web template if you know HTML well.
ii) Become a web designer - If you want to start a career as a professional web designer, HTML and CSS designing is a must skill.
iii) Understand web - If you want to optimize your website, to boost its speed and performance, it is good to know HTML to yield best results.
iv) Learn other languages - Once you understands the basic of HTML then other related technologies like javascript, php, or angular are become easier to understand.

Data mining and Data retrieval

Data Mining is the process of automatic discovery of novel and understandable models and patterns from large amounts of biological data. Bioinformatics is the science of storing, analyzing, and utilizing information from biological data such as sequences, molecules, gene expressions, and pathways. Development of novel data mining methods will play a fundamental role in understanding these rapidly expanding sources of biological data.

Over recent years the studies in proteomic, genomics and various other biological researches has generated an increasingly large amount of biological data. Drawing conclusions from this data requires sophisticated computational analysis in order to interpret the data. One of the most active areas of inferring structure and principles of biological datasets is the use of data mining to solve biological problems. Some typical examples of biological analysis performed by data mining involve protein structure prediction, gene classification, analysis of mutations in cancer and gene expressions. As biological data and research become ever more vast, it is important that the application of data mining progresses in order to continue the development of an active area of research within bioinformatics. This essay aims to draw information from varied academic sources in order to discuss an overview of data mining, bioinformatics, the application of data mining in bioinformatics and a conclusive summary.

Data mining

Data mining is the method extracting information for the use of learning patterns and models from large extensive datasets. Data mining itself involves the uses of machine learning, statistics, artificial intelligence, database sets, pattern recognition and visualisation. Often referred to as Knowledge Discovery in Databases (KDD) or Intelligent Data Analysis (IDA), the data mining process is not just limited to bioinformatics and is used in many differing industries to provide data intelligence. The application of data mining and machine learning models can involve varied systems.

This intelligence or knowledge discovery gained from data mining has a vast amount of aims, including the likes of forecasting, validation, diagnosis and simulations. Typically the process for knowledge discovery (see Figure 1) through databases includes the storing and processing of data, application of algorithms, visualisation/interpretation of results.

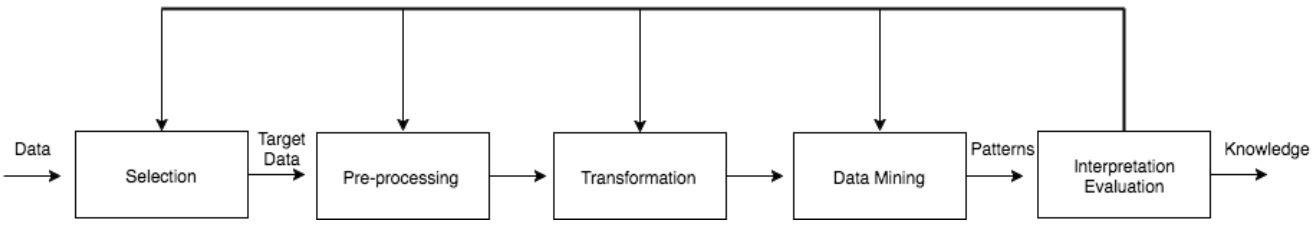


Figure: Process of Knowledge Discovery through Data Mining

It’s important to state that the process of data mining or KDD encompasses a multitude of techniques, such as machine learning. As a result the process of data mining includes many steps needed to be repeated and refined in order to provide accuracy and solutions within data analysis, meaning there is currently no standard framework of carrying out data mining.

Jain (2012) discusses that the main tasks for data mining are:

1. Classification: Classifies a data item to a predefined class

2. Estimation: Determining a value for unknown continuous variables

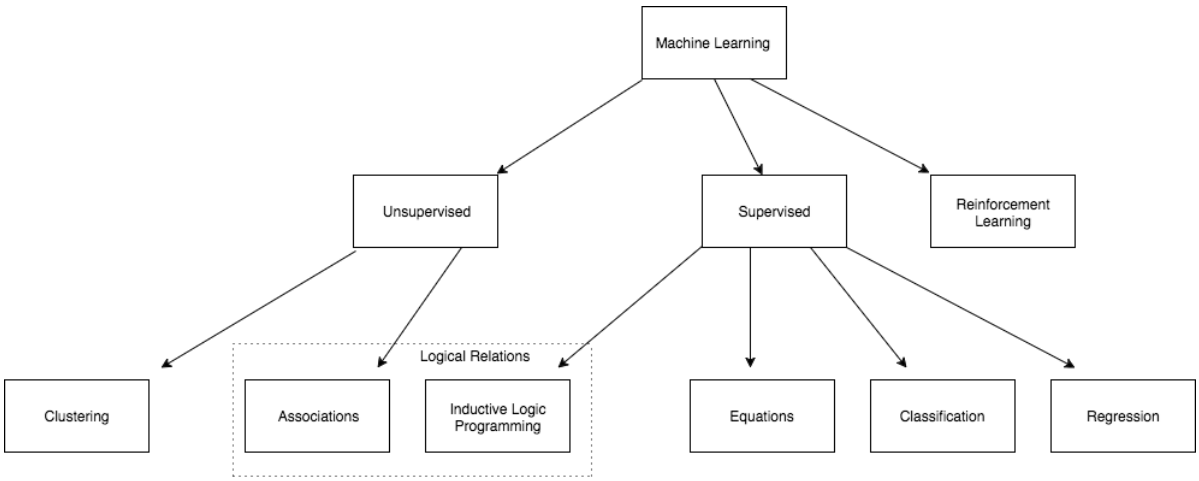
3. Prediction: Records classified according to estimated future behaviour

4. Association: Defining items that are together

5. Clustering: Defining a population into subgroups or clusters

6. Description & Visualisation: Representing data

Typically speaking, this process and the definition of Data Mining defines the extraction of knowledge. Where we define machine learning within data mining is the automatic data mining methods used. Following this, knowledge is gained through the use of differing machine learning methods used include: classification, regression, clustering, learning of associations, logical relations and equations.



Data mining (or knowledge discovery) is itself a multi-discipline area, involving machine learning, statistics, artificial intelligence, databases, pattern recognition, and data visualization. Generally, data mining consists of an iterative sequence of the following steps:

1) data cleaning: removing noise and inconsistent data from the original data;

2) data integration: combining multiple data sources consistently;

3) data selection: identifying and retrieving only the data that are relevant to the analysis task from the database;

4) data transformation: transforming and consolidating data into a format that is appropriate for mining by performing summary or aggregation operations;

5) data mining for knowledge discovery: applying intelligent machine learning methods to extract data patterns - this is the key process in knowledge discovery;

6) pattern evaluation: identifying interesting patterns that represent useful knowledge based on interestingness measures;

7) knowledge presentation: using intuitive visualization and effective knowledge representation techniques to present the discovered knowledge to the user.

Steps 1-4 are typically considered as data pre-processing steps for preparing the given data for the data mining task in Step 5. Steps 6 and 7 are often referred as decision support because they involve choosing interesting patterns/knowledge and presenting them to the user in a user-friendly manner for users to do further studies as well as decision making. Step 5, which is the key process for knowledge discovery, typically involves one or more of the following tasks:

• Association rule mining (Dependency modeling): This is invented and extensively studied by the data mining community. Its task is to detect relationships/associations between variables/features/items. One classic example is market basket analysis in which associations in supermarket customers' purchasing habits are mined from their sales transaction records ("market baskets") that can be translated useful insights for market campaigns. For example, the association rule {bread}=>{milk} can be mined in most supermarket data,indicating that if a customer has bought bread, he or she is likely to buy milk as well. With the association rules mined from the shoppers' transactional records, a supermarket/shop can automatically detect which products are frequently bought together and use this knowledge for marketing purposes, e.g., promotional bundle pricing or product placements. In the biomedical domain, association rule mining can be performed on gene expression data (or other medical data) to discover association rules where the antecedents are the biological features and their value ranges (cancer genes and corresponding gene expression values under different conditions; clinical test and corresponding readings/values) and the consequents are the class labels (cancer or non-cancer). The knowledge discovered can then be used for building a diagnostic system to assist doctors for decision making.

• Cluster analysis or clustering: This task's objective is to segment a set of objects into groups such that the objects within the same group are more similar to each other than compared to those in other groups. In machine learning, cluster analysis is a form of unsupervised learning, since there is no need for users or domain experts to provide training examples for clustering algorithms. One example use of cluster analysis in biomedical research is to segment gene expression data into groups where genes in each group share similar gene expression profiles, in order to discover genes that have the same biological functions.

• Classification: This task's goal is to assign the given input data into one of a known number of categories/classes. A classic example is spam filtering, in which the task is to classify a new email as a legitimate message or just spam. To build a classifier, a user

must first collect a set of training examples that are labeled with the predefined known classes (e.g., known spam messages and different types of legitimate email messages). A machine learning algorithm is then applied to the training data to build a classification model (classifier) that can be employed subsequently to assign the predefined classes to examples in a test set (for evaluation) or future instances (in practice). An application of classification in biomedical research is to predict the biological functions of novel proteins or genes. Here, proteins with known biological functions are first used as training examples to build a classification model which can be subsequently used to classify unknown proteins into one or more biological families with different functions. In this book, we also introduce the use of classification methods for predicting protein-protein interactions as well as drug-target interactions.
• Regression analysis: This task aims to find a mathematical function which models the data with the least error, where the focus is on the relationship between a dependent variable and one or more independent variables. Similar to classification, regression also requires training examples for building a regression function. In this case, each training example is associated with a numerical value instead of a class label as in the classification scenario. The difference between regression and classification is that regression handles numerical or continuous class attributes, whereas classification handles discrete or categorical class attributes. In this book, we have a chapter for trend analysis where the regression analysis is used as one of trend analysis techniques.

• Anomaly detection (Outlier detection): In this task, we attempt todetect data records/examples which do not conform to an expected or established normal behavior. The results could be interesting data records or erroneous records which require further investigation.

INFORMATION RETRIEVAL FROM BIOLOGICAL DATABASES

1. Entrez (NCBI)

Entrez is a retrieval system for searching several linked databases. It provides access to:PubMed : The biomedical literature; Genbank : Nucleotide sequence database Protein sequence database;Structure : three-dimensional macromolecular structures;Genome : complete genome assemblies; OMIM : Online Mendelian Inheritance in Man; Taxonomy : Organisms in GenBank

2. SRS (EBI and DDBJ)

SRS is a data retrieval system that integrates heterogeneous databanks in molecular biology and genome analysis. There are currently several dozen servers worldwide that provide access to over 300 different databanks via the World Wide Web. Additional technology to integrate externally developed applications into the package gives novel and powerful capabilities for biological data analysis.

3. DBGET/ LinkDB

DBGET is a data retrieval tool maintained by Kyoto University and the University of Tokyo. It covers more than 20 databases and is closely associated with the Kyoto Encyclopedia of Genes and Genomes. A related system LinkDB finds relationships between entries in the various databases covered by DBGET and others. DBGET has a simpler and more 'limited search format than Entrez.

SRS databases are grouped but use different principles to those used by Entrez and DBGET. For example, all sequences (nucleic acid and protein) are grouped together, while these are separated by Entrez. The use of SRS involves selecting one or more of these groupings and, within each selected group, selecting one or more of the available databases. Queries can be submitted using two styles of query form, standard or extended.